JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 7, Numbers 1/2, 2000
Mary Ann Liebert, Inc.
Pp. 215–231

DNA Segmentation Through the Bayesian Approach

V.E. RAMENSKY,[1] V.JU. MAKEEV,[1] M.A. ROYTBERG,[2] and V.G. TUMANYAN[1]

ABSTRACT

We present a new approach to DNA segmentation into compositionally homogeneous blocks. The Bayesian estimator, which is applicable for both short and long segments, is used to obtain the measure of homogeneity. An exact optimal segmentation is found via the dynamic programming technique. After completion of the segmentation procedure, the sequence composition on different scales can be analyzed with filtration of boundaries via the partition function approach.

Key words: nucleic acids, nucleotide composition, Bayesian statistics, segmentation.

1. INTRODUCTION: COMPOSITIONAL SEGMENTATION OF BIOLOGICAL SEQUENCES

Nucleotide sequences typically display correlations in nucleotide compositions, which are found both on small scales (Fickett, 1982; Frank and Makeev, 1997; Kypr and Mrazek, 1986; Mrazek and Kypr, 1994; Trifonov and Sussman, 1980; see also the literature on Markov chains, e.g., Guigo and Fickett, 1995) and on large scales (Sueoka, 1959; Bains, 1993; Bernardi, 1989; Bernardi, 1995; Bernardi *et al.*, 1985; Chechetkin and Lobzin, 1998; D'Onofrio *et al.*, 1991; Karlin and Brendel, 1993; Kypr and Mrazek, 1995; Kypr *et al.*, 1989; Li, 1994; Li, 1997; Ossadnik *et al.*, 1994; Peng *et al.*, 1992; Trifonov and Sussman, 1980; Tsonis *et al.*, 1991). Several biological issues are believed to be related to these correlations. Among other examples, one can mention the presumption of Bernardi *et al.* that many genomes of higher eukaryotes contain long regions of quasi-uniform composition, the isochores (Bernardi, 1989; Bernardi, 1995; D'Onofrio *et al.*, 1991), the compositional difference between exons and introns (Guigo and Fickett, 1995; Herzel and Grosse, 1997), simple repeats (e.g., Li and Kaneko, 1992), tracts in splicing sites (Gelfand, 1992; Gelfand, 1995), and DNA sites of factor binding (Gelfand and Koonin, 1997).

The local nucleotide composition is taken into account in many algorithms developed for the search of different patterns in DNA sequences (Krogh *et al.*, 1994(a); Krogh *et al.*, 1994(b); Reese *et al.*, 1997). Algorithms of this kind usually employ a fixed window, the length of which can affect the results. The preliminary compositional segmentation, which breaks a DNA sequence into homogeneous subsequences (blocks or segments), can therefore improve the perfomance of the pattern search.

Recently Bernaola-Galvan *et. al.* (1996) (see also Roman-Roldan *et al.* (1998)) have put forward a segmentation algorithm, where they used an approach based on the Jensen-Shannon divergence measure employed in various image-processing applications. The sequence is divided into two adjacent subsequences if the compositions of the resulting segments differ at the given level of statistical confidence. This procedure is repeated for each of the two resulting subsequences until the whole sequence becomes segmented

[1] Engelhardt Institute of Molecular Biology, Vavilova, 32 Moscow, 117984 Russia.
[2] Institute of Mathematical Problems of Biology, Institutskaya, 4 Puschino, Moscow region 142292 Russia.

into domains differing from each other in their composition. The segments so defined are neighborhood-dependent.

This technique, although conceptually simple and computationally efficient, has two basic disadvantages. First, regardless of the divergence measure chosen, the hierarchical segmentation procedure does not ensure that the true optimal segmentation will be attained. Indeed, the first boundary, which divides the sequence into two subsequences, and actually reflects the compositions averaged over very large distance, is retained for the whole subsequent procedure, even if it fails to be appropriate at the later stages of the segmentation process when it separates smaller blocks.

Second, this procedure requires a criterion for termination of cutting. To this end, Bernaola-Galvan *et al.* (1996) use statistical inference. The segmentation thus becomes dependent on the given threshold value indicating that the compositions of subsequences are significantly different. With the point estimator of the frequency of each letter used by Bernaola-Galvan *et al.* (1996), it is necessary that the segments are long enough to secure reliable statistics. This critical block length depends on the composition of the segments. This additional requirements may create difficulties in comparing the resulting segmentation with the functional structure of the sequence.

We believe that the optimal segmentation algorithm should be free of both shortcomings, with all possible configurations of boundaries taken into account and the short segments considered on an equal basis with the long ones.

Moreover, the very definition of homogeneous segments depends on the scoring system. Even such a simple sequence as `AAAACACACCCC` brings about the question of what segmentation, "`AAAA-CACA-CCCC`" or "`AAAACA-CACCCC`," is better. Different scoring systems can yield different answers.

Finally, the necessity to test all possible boundary configurations is a serious computational problem, since the number of possible segmentations is exponentially large (if a boundary is to be set between any two neighboring letters, this quantity equals 2^{N-1}, where N is the sequence length). The computational time may be reduced with scores of a special type, for which dynamic programming or another similar technique may be employed.

In the context described, the Bayesian approach turns out to be highly expedient. For the problem discussed, it is intuitively very clear. All the integrals involved in constructing the Bayesian estimators can be taken analytically, which is a rare case in the practical application of the Bayesian statistics. The Bayesian estimator does not degenerate for short blocks, and one can easily incorporate it into the dynamic programming scheme.

In the framework of general Bayesian inference, this problem relates to the multiple-changepoint identification (Chernoff and Zacks, 1964; Smith, 1975; Booth and Smith, 1982). The general theory, with examples, is presented by Stephens (1994). Segmentations of sequences with the Bayesian approach was recently discussed by Liu and Lawrence (1999) (see also Lawrence (1997)). Their paper can also serve as an excellent introduction to the Bayesian approach in bioinformatics. Basically, their technique is free of the disadvantages of the approach of Ramon-Roldan *et al.* and begins with segmentation of the sequence into a fixed number of blocks, with subsequent calculation of the boundary probability distribution using a sampling procedure (see also Stephens (1994)). We discuss the differences of this algorithm with ours in detail in Section 6. The only remark to be made in advance is that we use sets of possible boundaries rather than probability distributions for borders and do not use sampling.

Our algorithm consists of the following two stages: (i) given the sequence, the possible configurations of boundaries between blocks, i.e., the segmentations, are tested. For each segmentation a likelihood function is calculated. This function is maximized over all possible segmentations via dynamic programming, yielding the optimal segmentation. As a rule, the optimal segments are relatively short. Therefore, to obtain segmentation with larger segment lengths, some neighboring blocks of similar composition should be merged, that is, some boundaries should be filtered out.

For filtration (ii), we calculate the probability of every optimal boundary. To this end, we use the partition function of the whole sequence, as well as the partition functions of the subsequences to the right and to the left of the boundary under evaluation. After filtration, only boundaries with probabilities higher than a given threshold are retained, providing the (suboptimal) segmentation with longer segments.

In this paper we also discuss the practical aspects, such as the choice of the prior, the efficient algorithm, the examples of segmentation of biological sequences, and the possible future applications of the method.

2. BAYESIAN SEGMENTATION. GENERAL CONCEPT AND CONSTRUCTION OF ESTIMATORS

2.1. *The problem*

Consider a symbolic sequence written in an alphabet Ω containing L letters. We are going to approximate this sequence with a probabilistic model. The sequence is considered as a series of statistically independent and compositionally homogeneous blocks of arbitrary length. These blocks are of multinomial random nature. For each block, we shall refer to the observed numbers of each letter in it as to the counts $\mathbf{n} = (n_1, \ldots n_L)$. These counts serve as arguments of a score function that reflects the homogeneity of each block. We define a segmentation as a set of boundaries separating these blocks and use block counts to estimate their compositions. Since the blocks are assumed to be statistically independent, the score for the entire segmented sequence is a product of the scores for individual blocks.

With the score function at hand, the sequence is studied with two complementary techniques, whose combination yields reliable results. The result of applying the first technique is a single optimal segmentation with the maximal score; the second technique consists in calculating the probabilities of boundaries with the help of the partition function of all possible segmentations.

2.2. *Optimal segmentation*

To obtain the optimal segmentation, we employ a dynamic programming algorithm (Finkelstein and Roytberg, 1993), which computes it for the time proportional to the squared sequence length (see Section 4). However, several various partitions often yield the same score (e.g., `AACAC` and `A-ACAC` for the marginal likelihood score; see Section 2.7) and our algorithm provides only one solution out of the many having identical scores.

Moreover, some of the boundaries forming the optimal segmentation are very sensitive to minor changes in the sequence. Although the scores of the optimal segmentation of the changed sequences are close to the score of the unperturbed one, the segmentation patterns can be rather different.

The second technique can take into account the suboptimal solutions.

Boundary probabilities. Partition function. Determine the partition function of the set of segmentations of a sequence of length N in a standard way (Finkelstein and Roytberg, 1993) by summing the probabilities of all possible partitions:

$$Z(N) = \sum_{q_1} \cdots \sum_{q_{N-1}} \Pi(q_1, \ldots q_{N-1}) \tag{1}$$

where the indicator q_k equals 1 if there is a boundary located after the letter k in the sequence, and 0 otherwise; the overall $\mathbf{q} = (q_1, \ldots, q_N)$ determines a segmentation which has the probability $\Pi(\mathbf{q})$. The partition function can be easily computed via the dynamic programming algorithm (see 4.2 below).

To calculate the probability of the boundary located after the letter k, one needs two partition functions of subsequences to the left and to the right of this border, ZL and ZR respectively:

$$\Pi(q_k = 1) = \frac{ZL(k)ZR(N-k)}{Z(N)}. \tag{2}$$

Boundaries with higher probabilities separate blocks with higher compositional divergence. Therefore, such boundaries tend to be more stable with the sequence undergoing minor changes. The partition function approach cannot provide an optimal segmentation of the sequence, but can indicate which boundaries are more statistically significant. Thus, the two techniques complement each other in obtaining a stable suboptimal segmentation pattern.

In the following sections we obtain the score that reflects the block homogeneity.

2.3. *Blocks of random composition. Probabilistic approach*

Denote the set of letter probabilities of a multinomial (Bernoulli) sequence as $\sigma = \{\theta_1, \theta_2, \dots, \theta_L\}$, which is subject to the normalization condition

$$\sum_{k=1}^{L} \theta_k = 1. \tag{3}$$

We shall use σ to represent the composition of the block. The objective is to find compositions which provide a high probability of obtaining the block sequence in a series of random tests.

Given the composition, the probability of the counts is $P(\mathbf{n}|\sigma) = N!\prod_{i=1}^{L}[\theta_i^{n_i}/n_i!]$, whereas the likelihood of the individual sequence to occur is

$$P(S|\sigma) = \prod_{i=1}^{L} \theta_i^{n}. \tag{4}$$

Adoption of the simplest frequency-count estimator of the composition $\hat{\theta}_i = n_i/N$ results in the trivial segmentation down to the individual letters which have 100% probability; this is certainly not what we need. Moreover, the trivial segmentation is statistically unreliable. Thus, the size of the sample should be taken into consideration, especially for very small samples. This is exactly the case for which the Bayesian techniques excel. The Bayes estimator can be viewed as a way to directly account for the "missing information" when the information at hand consists of a finite data set rather than a full distribution $p(\sigma)$ (Wolpert and Wolf, 1995).

2.4. *Composition probability density function*

The Bayesian statistics regards all the parameters of the problem as random variables. In the beginning of the procedure, these variables have some prior probability distribution, which can be chosen quite arbitrarily. The experimental data allow one to estimate these probability distributions via the Bayes formula. The result of the Bayesian estimation is again some probability distribution of the estimated quantity, rather than a single value considered in the classical statistics (Rozanov, 1985).

However, Bayesian and classical statistics agree for the large samples. More precisely, if we estimate parameters from large samples, the Bayesian probability distributions converge with probability 1 to the correct parameter values for any prior (Rozanov, 1985). Thus, the Bayesian estimators are consistent. However, the choice of prior is much more important for the small quantities of data, which is exactly our case, since we also consider very short blocks.

The results of previous estimations may be used as the prior for the next step. This technique is referred to as recurrent estimation and is employed, for instance, in adaptive learning algorithms.

For the composition $\sigma = \{\theta_1, \theta_2, \dots, \theta_L\}$, write the probability density function $p(\sigma)$, which is a function defined on the simplex $\Psi = \{\sigma : \theta_k \geq 0; \sum_{k=1}^{L} \theta_k = 1\}$ satisfying the normalization condition:

$$\int_{\Psi} d\sigma p(\sigma) = 1. \tag{5}$$

The integral above may be understood in two senses. Wolpert and Wolf regard it as an ordinary L-dimensional integral over the region extending in each dimension from 0 to ∞, whereas the function under integration has support only on the simplex Ψ. The other point of view assumes it to be a first-type surface integral taken over the surface of the simplex. These two cases are different for the $\sqrt{L!}$ multiplier related to the total surface of the simplex Ψ, which is $\frac{\sqrt{L!}}{(L-1)!}$ (see normalization (3)). In our case, this multiplier cancels in final equations, which thus do not depend on the form of the integral. Further than that, we shall not use this multiplier.

Given some prior distribution $p(\sigma)$, consider a putative segment with sequence S. The Bayes theorem brings about the posterior probability density function $p(\sigma|S)$

$$p(\sigma|S) = \frac{P(S|\sigma)p(\sigma)}{P(S)} \tag{6}$$

where

$$P(S) = \int d\sigma P(S|\sigma)p(\sigma) \tag{7}$$

is the normalization constant called the marginal likelihood (Stephens, 1994; Liu and Lawrence, 1998). Note, that it is obvious from (4) that the marginal likelihood depends only on the counts **n** and not on the order of letters in sequence **S**. In this paper, we basically use the notation of Wolpert and Wolf (1995), which agrees with a statistical physicist's intuition. However, we have changed some of their notations to those more common in computational biology.

2.5. Prior distributions. Informative versus noninformative priors

Since our analysis includes very short blocks as well, the character of the prior distribution is very important. The success of the Bayesian techniques depends dramatically on the prior used, especially for small samples, and the proper choice of the prior should be conditioned by the context of the problem, as there are no formal recipes.

If the prior that the researcher adopts does not match the prior generating the data, the Bayesian estimators are far from optimal. The Bayesian approaches are "only as good as the prior." Other statistical techniques, such as maximum likelihood, have the advantage that their predictions do not depend on the assumption of a prior. On the other hand, Bayesian statistics is very convenient in optimization problems, where the probability distribution is also the subject of optimization, such as in adaptive methods. Another advantage of Bayesian techniques is that they make all assumptions explicit, by putting them in the prior (Wolpert and Wolf, 1995).

The simplest prior is a uniform distribution, or noninformative prior (Dunbrack and Cohen, 1997). In many problems, priors other than uniform must be used. Several computationally effective priors are known in the Bayesian statistics, such as the Dirichlet prior and the entropic prior (see the discussions in Wolf and Wolpert (1993, 1995) and Liu and Lawrence (1998).

For the segmentation problem, the choice of the prior actually implies some ideas about the overall composition of a polymer. In practice, some statistical observations are made on data banks, or on the composition of the same sequence averaged on a larger scale. Thus, some correlation of the sequence composition with other sequences or with other parts of the studied sequence is introduced into the inference. This is particularly evident when the Dirichlet prior is used (Liu and Lawrence, 1998), the very form of which exhibits counts made beforehand. The more complex entropic prior (Wolf and Wolpert, 1993, 1995) reflects the statistical homogeneity of the prior source data.

In other words, the choice of any informative prior means that some "natural" statistical dependence between the block compositions is expected, and the blocks appear as members of a greater family of sequences with determined statistical properties. But this is essentially the subject of the experimental (statistical) testing. It is unjustified to make any guess on the sequence composition before studying the statistical pattern of the whole data bank. In our opinion, the overall data bank of sequences and even all parts of a sufficiently long sequence, especially eukaryotic one, make a complex mixture (Sjolander, 1996), the components of which are not very stable with small changes of the data (see D'Onofrio *et al.*, (1991) for the experimental evidence and Li (1997) for review). Therefore, the best choice for the initial studies appears to be a noninformative prior, which does not suggest any statistical dependence.

Observations on data banks (Bernardi *et al.*, 1995; Li, 1997) indicate that probably some clustering in the composition of different sequences take place, and some values of letter frequencies seem to be preferred. This situation is reflected by the Dirichlet mixture prior. We believe that this kind of prior is likely to improve the results. However, the parameters of the mixture are to be obtained experimentally, and the technique we use is one of the ways of studying the correlations necessary to construct a better mixture prior.

2.6. Marginal likelihood. Recurrent estimation

Inspecting (7), one can see that the marginal likelihood is the probability of obtaining the test Bernoulli-type sequence in the experiment with the frequencies σ picked up from the ensemble of uniform composition (for the uniform prior). In other words, we pick the composition and then start a Bernoulli series of N tests. But this is exactly the context in which our problem is formulated: we assume that DNA is a set of segments having unknown but fixed compositions and try to find the approximation that agrees best with the data (the sequence).

It is easy to calculate the marginal likelihood (7) for the uniform prior:

$$P(S) = P(\mathbf{n}) = \frac{(L-1)!}{(N+L-1)!} n_1! \ldots n_L! \tag{8}$$

(Liu and Lawrence, 1999).

Integration is readily performed using the following expression:

$$\int_{\Psi} d\theta_1 \ldots d\theta_L \theta_1^{n_1} \ldots \theta_L^{n_L} = \frac{n_1! \ldots n_L!}{(n_1 + \ldots + n_L + L - 1)!}. \tag{9}$$

This integral may be calculated with different tricks, such as the Laplace transform (Wolpert and Wolf, 1995) or change of variables (Grosse, 1996).

Expression (8) also may be obtained by recurrent estimation. Take the first letter from the putative block. It is considered when no information about the composition of the putative block is available (for the uniform prior). The probability of finding each letter is $P(1, 0, \ldots, 0) = \int_{\Psi} d\sigma \theta p(\sigma) = 1/L$, which looks reasonable. Then we can re-estimate the probability density function: $p(\sigma|1, 0, 0, 0) = L!\theta_1$. For the four-letter DNA alphabet, it is easy to check that the probability to obtain the same letter in the next experiment will be $2/5$, and the probability to obtain another letter will be $1/5$. The formula of conditional probability $P(AB) = P(A)P(B|A)$ should be used to obtain the total probability of a two-letter sequence. In the case of the four-letter DNA alphabet, $P(AA) = 1/4 \cdot 2/5 = 1/10 > P(A)P(A) = 1/16$ and $P(AC) = 1/4 \cdot 1/5 = 1/20 < P(A) \cdot P(C) = 1/16$. Therefore, this estimator, like the extreme likelihood score, does not degenerate for the shorter blocks.

In the general case, the probability to obtain the i-th letter knowing that the previously obtained sequence has counts $\mathbf{n}$ can be written as

$$P(i|\mathbf{n}) = \frac{n_i + 1}{N + L}, \tag{10}$$

(this is the classical Laplace sample size correction estimator, which can be easily calculated directly from (6) (see Dunbrack and Cohen, 1997; Grosse, 1996; Wolf and Wolpert, 1993). For sequences of arbitrary length, direct integration demonstrates that the estimation procedure depends not on the order of letters, but on the counts $\mathbf{n}$. The overall probability of the sequence can be calculated with the use of the conditional probability formula multiplying estimates (10) for each observed letter up to the corresponding N, starting with $1/L = \frac{(L-1)!}{L!}$. The result of recurrent estimation coincides with the marginal likelihood (8). By definition, it is a normalized probability determined on a set of sequences of fixed length N.

We refer to this weight as the marginal likelihood weight.

Liu and Lawrence (1998) obtained this weight in a more formal manner by considering the segmented DNA sequence as a subject for direct Bayesian estimation of both the boundary positions and the block compositions. Thus, they had to introduce an additional prior of segment boundaries located at the positions between particular letters.

However, in the case of long segments, the marginal likelihood score does not provide a large number limit consistent with the segmentation measures obtained from the informational theory approach (see appendix). Thus, in order to compare our segmentation results with those obtained from informational theory, we need another homogeneity measure, which corresponds to the information divergence for large blocks. This measure is discussed in the next section.

2.7. Extreme likelihood

Another way to obtain the weight of a segment is to average the likelihood function (4) with the estimated probability density (6). We shall refer to this score as the extreme likelihood score. It characterizes the likelihood of the putative block generation in the most favorable case when the composition is estimated from the putative block itself. For the noninformative prior $p(\sigma) = (L-1)!$ the probability density is

$$p(\sigma|\mathbf{n}) = \frac{(N+L-1)!}{n_1! \ldots n_L!} \theta_1^{n_1} \ldots \theta_L^{n_L}. \tag{11}$$

With this probability density, the extreme likelihood is

$$L^{++}(\mathbf{n}) = \frac{(N+L-1)!}{n_1!\dots n_L!}\frac{(2n_1)!\dots(2n_L)!}{(2N+L-1)!}. \tag{12}$$

The estimator (12) does not degenerate for single-letter blocks. Indeed, for instance, for the four-letter DNA alphabet $L^{++}(A) = 2/5$; $L^{++}(AA) = (5!4!)/(2!7!) > 4/25$; $L^{++}(AC) = 2/21 < 4/25$. Furthermore we shall show that it is a good limit for large N. However, it has some disadvantages. First of all, it tends to overestimate the probabilities of the shorter segments. For instance, the (ACG)$_\mathrm{n}$ sequence is naturally expected to be a single block. Closer inspection (using the Stirling formula), however, demonstrates that given this weight, the segmentation of this sequence to individual letters is the best for any N. The second drawback is that the block weights cannot serve as normalized probabilities for all sequences of fixed length: for instance, for $L = 4, N = 1, L^{++}(A)+L^{++}(C)+LL^{++}(G)+L^{++}(T) > 1$.

3. LONG SEGMENTS. COMPARISON WITH INFORMATIONAL APPROACH

We are going to show now that in the case of long segments our results yield the segmentation that is consistent with the one obtained via the informational technique used, for instance, by Roman-Roldan *et al.* (1998) and Bernaola-Galvan *et al.* (1996). When all components of $\mathbf{n}$ are sufficiently large, the Stirling approximation $n! = \sqrt{2\pi n}n^n e^{-n}$ may be used. The extreme likelihood weight thus approximates (see appendix)

$$L^{++}(\mathbf{n}) = \frac{1}{2^{(L-1)/2}}e^{-NI(\mathbf{n})} \tag{13}$$

where $I(\mathbf{n}) = -\sum_{l=1}^{L}\hat{\theta}_l \ln\hat{\theta}_l$ is the information content (multipied by $\ln 2$) and $\hat{\theta}_l = \frac{n_l}{N}$ is an estimator for the letter i content; frequency count and Bayesian estimations yield close values in the case of large N (compare with (10)).

Thus, one can see that the segmentation that uses an extreme likelihood score is essentially an entropic segmentation for large N and may be considered as an extension of the entropic segmentation for the case of small N. It should be stressed here that if we consider a direct Bayes estimator for the entropy (Grosse, 1996; Wolf and Wolpert, 1993, 1995), the closed form of the result will be different from (12). The question arises whether it is more statistically natural to consider estimators that are reasonable at small lengths and provide a good large-distance limit, or to try to build estimators of the closed-form entropy function on small samples. The results of our preliminary computations (Ramensky *et al.*, 1998) indicate that the Grassberger estimator for entropy (Grosse, 1996) yields contents even more overly estimated than the extreme likelihood weight, thus providing segmentation down to very short blocks. However, this estimator is known to overestimate the entropy itself.

In the procedure used by Bernaola-Galvan *et al.* (1996) (see also Roman-Roldan *et al.*, 1998), the sequence is divided into two parts of lengths N and M minimizing the Jensen-Shannon divergence measure (Lin, 1991): $I_{JS}(\mathbf{n},\mathbf{m}) = I(\mathbf{n}+\mathbf{m}) - \frac{N}{N+M}I(\mathbf{n}) - \frac{M}{N+M}I(\mathbf{m})$, where $\mathbf{n}, \mathbf{m}, \mathbf{n}+\mathbf{m}$ are vectors of the counts of sequence parts and the sequence whole, respectively. This term appears in the expression for the likelihood ratio function of the partition in the form

$$\frac{L^{++}(x_1,\dots,x_N)L^{++}(y_1,\dots,y_M)}{L^{++}(x_1,\dots,x_N,y_1,\dots,y_M)} \xrightarrow[N,M\to\infty]{} \frac{1}{2^{(L-1)/2}}e^{-(N+M)I_{JS}(\mathbf{n},\mathbf{m})} \tag{14}$$

Therefore, our technique incorporates the approach of Bernaola-Galvan *et al.*, providing a good small-distance weight.

Roman-Roldan *et al.* (1998) build a measure of sequence complexity using the function which is subject to minimization in our Section 2.2.

The marginal likelihood estimator for the large N approximates

$$P(\mathbf{n}) = \left(\frac{2\pi}{N}\right)^{(L-1)/2}\hat{\theta}_1^{n_1+1/2}\dots\hat{\theta}_L^{n_L+1/2} = \left(\frac{2\pi}{N}\right)^{(L-1)/2}\sqrt{\hat{\theta}_1\dots\hat{\theta}_L}e^{-NI(\mathbf{n})}. \tag{15}$$

For the problem of partition of a sequence into two segments, this weight yields for large N and M the following likelihood ratio function of the partition:

$$\frac{P(x_1, \dots, x_N)P(y_1, \dots, y_M)}{P(x_1, \dots, x_M, y_1, \dots, y_M)} = \left(\frac{2\pi NM}{N+M}\right)^{(L-1)/2} \sqrt{\frac{\hat{\theta}_1^{(N+M)} \dots \hat{\theta}_L^{(N+M)}}{\hat{\theta}_1^{(N)} \dots \hat{\theta}_L^{(N)} \hat{\theta}_1^{(M)} \dots \hat{\theta}_L^{(M)}}} e^{-(N+M)I_{JS}(\mathbf{n},\mathbf{m})}. \tag{16}$$

The Jensen-Shannon divergence measure also appears here in the exponential function, but the pre-exponential multipliers are also dependent on **n**.

4. ALGORITHM. DYNAMIC PROGRAMMING

Given the score function, the algorithmic part of the segmentation problem is formulated as follows. Consider the sequence $S = s_1 s_2 s_3 \dots s_N$ of length N, where $s_i \in \Omega$. For every segment $S(a, b) = s_a \dots s_b$, $a \in b$, there is weight $W(a, b)$, which equals $\ln P(S(a, b))$ for the marginal likelihood score, and $\ln L^{++}(S(a, b))$ for the extreme likelihood score.

Any particular segmentation R in m blocks is determined as a set of boundaries $R = \{k_0 = 0, k_1, \dots, k_{m-1}, k_m = N\}$, where k_i separates s_{k_i} and s_{k_i+1}; $k_0 < k_1 < \dots < k_{m-1} < k_m$. Define the weight of segmentation R as

$$F(R) = \sum_{j=1}^{m} W(k_{j-1} + 1, k_j). \tag{17}$$

4.1. *Optimal segmentation*

The optimal segmentation R^* that maximizes $F(R)$ is found in the recurrent manner (Finkelstein and Roytberg, 1993). Denote by $R^*(k)$ the optimal segmentation of the fragment $S(1, k)$, $1 \leq k \leq N$. Calculation of $R^*(1)$ is trivial. In the case of known optimal segmentations $R^*(1), \dots, R^*(k-1)$, the optimal segmentation $R^*(k)$ is found using the following recurrent expression

$$F(R^*(k)) = \max_{i=0,\dots k-1} [F(R^*(i)) + W(i+1, k)] \tag{18}$$

Here, we put $F(R^*(0)) = 0$. Since the building of a segmentation $R^*(k)$ takes the time $\sim k$, the total time can be estimated as N^2.

4.2. *Partition function calculation*

For functions determined on the segmentations, we shall also use another set of variables, namely the indicators of the boundary positions q_k, $k = 1, \dots, N$. By definition, $q_k = 1$ if there exists a segment boundary after the letter k, otherwise $q_k = 0$. We shall use both variables, $F(R)$ and $F(q_1, \dots, q_k)$, without special comments.

Using (17), the partition function (1) may be rewritten as follows (Finkelstein and Roytberg, 1993):

$$Z(N) = \sum_{q_1} \dots \sum_{q_{N-1}} e^{F(q_1, \dots q_{n-1})}. \tag{19}$$

To calculate the probability of a boundary after the letter k, we need also the partition functions of the segments to the left and to the right of this boundary:

$$ZL(k) = \sum_{q_1} \dots \sum_{q_{k-1}} e^{F(q_1, \dots, q_{k-1})} \tag{20}$$

$$ZR(k) = \sum_{q_k} \dots \sum_{q_{N-1}} e^{F(q_k, \dots, q_{N-1})}. \tag{21}$$

The recurrent formulae to calculate $ZL(k)$ and $ZR(k)$ are analogous to (18) and are obtained through the formal substitution of operations. Summation is used instead of taking the maximum, and multiplication is used instead of summation (Finkelstein and Roytberg, 1993). Then the following relations correspond to (18):

$$ZL(k) = \sum_{j=0}^{k-1} e^{W(j+1,k-1)} ZL(j) \tag{22}$$

$$ZR(k) = \sum_{j=k}^{N} e^{W(k,j)} ZR(j) \tag{23}$$

with the respective boundary conditions

$$ZL(0) = ZR(N+1) = 1; \quad W(k-1,k) = W(N,N+1) = 0. \tag{24}$$

Note that this problem can be reformulated as an optimal path problem for a corresponding weighted oriented acyclic graph (Finkelstein and Roytberg, 1993; cf. also the Viterbi algorithm, Ryan and Nudd, 1993). The graph G has $n+1$ vertices, or states $V_0, \ldots, V_\mathrm{n}$ in terms of Viterbi, where a vertex V_i corresponds to the boundary between the ith and $(i+1)$st symbols of the given sequence. The edges connect all pairs of vertices $(V_\mathrm{i}, V_\mathrm{j})$, where $i < j$. The weight of the edge $V_\mathrm{i}V_\mathrm{j}$ is equal to $\log W_{\mathrm{i}+1,\mathrm{j}}$. It is obvious that the path from V_0 to V_n having a maximal possible weight corresponds to an optimal segmentation.

Relations (22, 23) can be obtained straightforwardly, without substitution of operations. To prove (22), divide all partitions of the initial segment $S(1,k)$ into groups according to the last block in the partition. Then each term in the sum (22) corresponds to such a group.

5. EXAMPLES OF SEGMENTATIONS OF GENOMIC SEQUENCES

In this section we discuss the implementations of the algorithm described and correlate the computed change points with the textual patterns found in genomic sequences with $L = 4$. The first two examples illustrate the patterns found in the optimal segmentation without any filtration. In all examples, we use the marginal likelihood score, which is more consistent with the general Bayesian framework and definitely works better when the segments are short.

Several preliminary remarks should be made. When the same letters neighbor, they always belong to one segment. This results from the fact that the log-score is an additive function. The same is true for dimers such as ACAC. On the other hand, large number statistics studied in Stirling approximations demonstrates that long strands containing only one letter or equal shares of two or three letters belong to one block. Conversely, a long strand containing equal shares of all four letters has a score lower than that obtained for the individual letters.

This is the result of adopting a uniform prior. If one uses an informative prior with significantly biased expected frequencies, then the sequence containing equal shares of all four letters can preferably belong to one block.

The arguments above are illustrated with the first example (Fig. 1), which is the optimal segmentation of the first 1000bp of yeast chromosome I. One can see that the regions with dramatically biased compositions constitute individual blocks (shown with bold letters). These are the telomeric region containing only A and C bases, and another 141bp region starting at position 154 and containing only four occurrences of G.

At the same time, regions with more uniform composition tend to be segmented into much shorter blocks containing only two or three types of letters. There are also plenty of single-letter "segments."

Short segments containing not all letters of the alphabet can also be of biological interest. This is illustrated with Figure 2, which presents the optimal segmentation of human promoter IGFBP4 (EMBL sequence HSIGBP4, (Zazze *et al.*, 1998)). One can easily see that most segments observed are the palindromic (or similar) sequences abundant in eukaryotic promoters.

ccacaccacacccacacacccacacaccacaccacacaccacaccacacccacacacacaca t cc t aaca c t a ccc t aaca g

ccc t aa tct aa ccc t gg cc aa cc t g tctctc aa c tt a ccctcc a tt a ccc t g cctccactc g tt a

ccc tgt **cccataaccataccactccgaaccaccatccatccctctacttactaccactcacccaccgttaccctccaattacccatatcaccc>**

<actgccacttaccctaccattaccctaccatccaccatgacctactcaccatac t g ttcttct a cccacc atataaa cgc t

aacaaa t g a t c g t aaataa cacacac gtg c tt a ccc t a ccac tttat accaccaccaca t g

ccatactcactcac tt g tata c t g a tttt a c g t a c g cacac gg a t g c t aca g tatata cc a tctc

aaa c tt a ccc t a ctctc aga tt cc tt c a ctcc a t gg ccc a tctctc a c t g aa t c a g t a cc aaa

t g cactcacatca ttat g cac gg cac tt g cctc a gcgg tata ccc tgtg cc a ttt a ccc ataa cgccc a t c

a ttat ccac a tttt g atatctatatctcatt c ggcgg t ccc aaa ttgt ataa c t g ccc tt aatacata c g ttat

a ccac tttt g cacc atata c tt a ccactcc atttatata cac ttat g t c tatt a c agaaaaa t ccccac aaaaa

t cacc t aaacataaaaata ttct a c tttt c aacaataatacataaacata ttgttgtggt a g caacac tat c a t gg

tat cac t aa c g t aaaa g tt cctc aa tatt g c aa ttt g c tt g aa c gg a t g c tattt g aata ttt c

g t a c tt acaca gg cc ata c attagaataatat g t cacatcac tgt c g t aaca ctc tttatt cacc gag c t

aata c gg t a gtgg ctc aaa ctc a t gcggg t g c tat g a t acaa ttatatcttattt cc a tt ccc atat g c

t aa ccgc aa t cc t aaaa g c ataa c t g a t g c a tcttt aa t c ttgtatgt g acactactcatac g aa ggg

a c tatat c t a g t c aaga c t a c tgtg ata gg t a c g ttatttaata gg a tct ataa c g aaa tgt c

aaataa tttt a c gg taatataa c ttat c a gcgg tata c t aaaa c gg a c g tt a c g ata ttgt

FIG. 1. Optimal segmentation of the first 1000bp of yeast chromosome **I**.

```
cctc aaa t cc tt cac t g cc a tct ggg aaa t g a t cacaaca g ctct acaaatacacaa t g a tt acaa gg
aa t ggtg ccccac t ggag ttgtt c aa cgc aaaa c tt g caca tt g c aa gtgg c aa t ctccc a gg cc t g
cctccctcc a c g a gtggg tct g aa t ggg cc t gagagg c aaaca t cc aagaaggaggaagagg c t cggcggc a
cctccctcccc gggag ttct g c t g a tt cc a tctt gggg aa g c a gggtgg a cc a ggg ccc aaa t g cgccc
t ggggag a tt gcgggggcgggagagg tt g c aa gggg c aa gtgg c aaga g cc tgtt aa c g tctt a ggg cctcc
a gg cc tttct gtg cccc t a g c tgtg cc tgt a cgc ttt a ccccacc t c a ggagg c tt gg tctcc a gcgg
tt gagg c t gg aa g cacc ggggtgcggtgg aaa ggg ctct g t cc aggaaga cc gg a t ccgc agag cc gggag t
cc ggg c t a gg aa g tccctttctc ggtggg aga c t gagg ccgcc tt ggcggggcggg a c gaga ctcctcc gagg t
c ggg aaa ggggg ccccgc a g c a g cccc tt gg cttcccttctcccttgcctcccctcc gggg ctcc gg tt c a gagg
c a ctct ggg cgcc t g c t aca g c tt cc aaa c t gcgccgc ttccttcttc gg c agaaaagga c ttt c aga t
gcggcggcggcggcggcggcg a ctc a gg aca gcg ccccctcccc t aa c gg
ccgcctctccctctccccctcgcccgccccggctccccc a cctc t ggg aa ggcg c t ggggtgtgg cc a ggg a cc gg t
ataaa g t cc gggggag cc gg t ccc ggg
```

FIG. 2. Optimal segmentation of the human **IGFBP4** promoter.

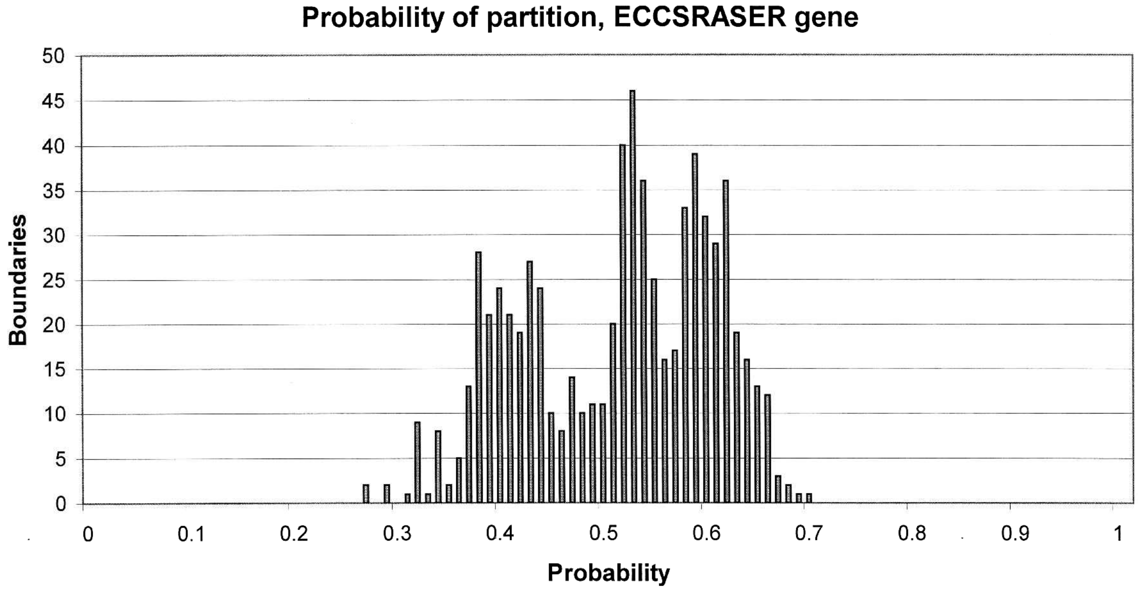

FIG. 3. Histogram of partition probabilities of the *E.coli* **ECCSRASER** region.

If one is interested in segmentation on a larger scale, some segments with close compositions should be merged. In our algorithm, this is done with the help of the partition function.

Generally, the boundaries between long blocks with different compositions correspond to the partition function maxima. However, this probability is a rather smooth function of the sequence position (see discussion in Section 6), and the inner boundaries of compositionally biased segments close to their margins also have relatively high values. Thus, the partition function probabilities should be used with caution.

In our approach, we calculate the optimal segmentation and then use the partition function to merge blocks produced by the optimal segmentation. In so doing, the partition function is calculated taking into account only the segmentations obtained by merging the optimal blocks. In this case, the inner boundaries of compositionally biased segments do not obscure the picture.

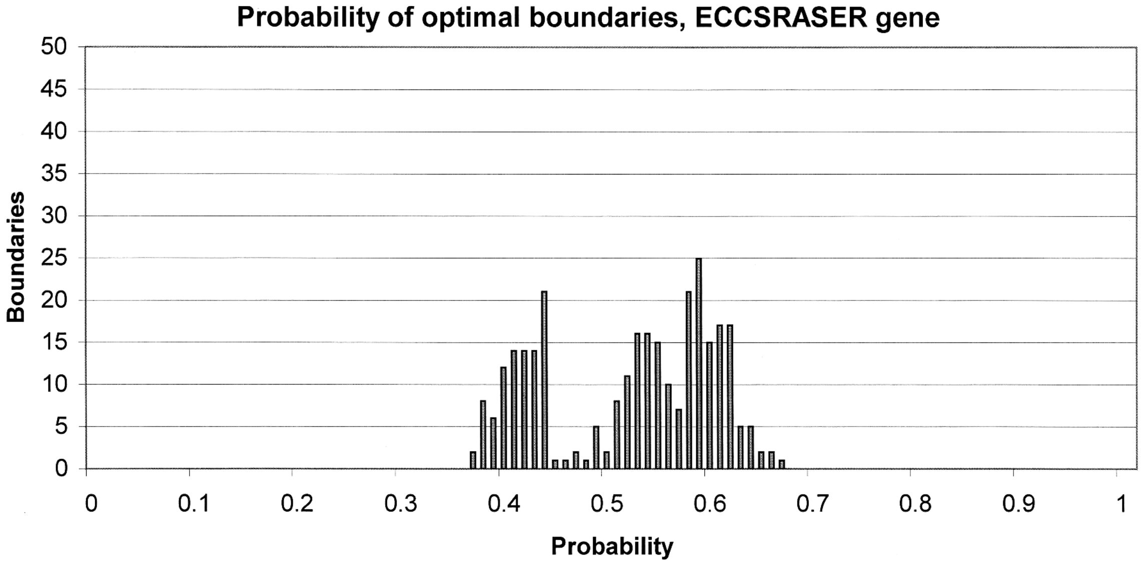

FIG. 4. Histogram of partition probabilities for the boundaries of optimal segmentation of the *E. coli* **ECCSRASER** region.

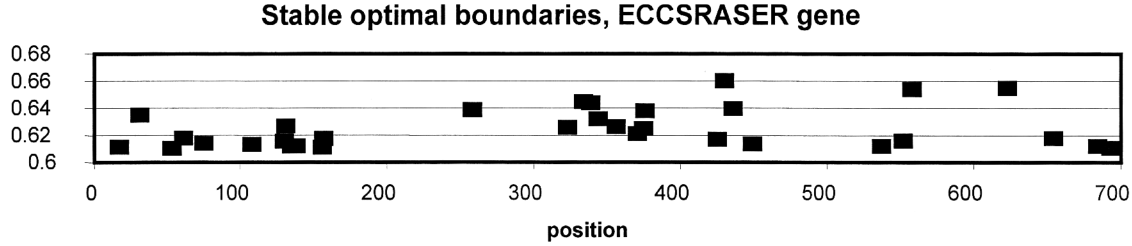

FIG. 5. Filtered boundaries of optimal segmentation of the *E.coli* **ECCRASER** region.

This is illustrated with Figures 3 and 4. Figure 3 displays the histogram of the partition probabilities of segmentation of the *E. coli* 708bp sequence region coding the ECCSRASER gene. The optimal segmentation consists of a number of quite short blocks. Figure 4 displays the histogram of boundary probabilities resulting from the optimal segmentation. One can clearly see that modes in Figure 4 are more distinct than in Figure 3, which is the result of excluding the inner boundaries of optimal segments. Boundaries corresponding to the points in the right part of the histograms separate long regions of different composition. Figure 5 shows the boundaries with probabilities (in Figure 4) higher than 0.61. These segments correlate with the coding regions and intergenic elements.

6. DISCUSSION

Our segmentation algorithm is free of heuristic approximations and provides a reasonably large N limit. Both optimal segmentation and partition function calculations require a computational time of the order of squared sequence length. In this section, we are going to compare our technique with that of Liu and Lawrence, who use a similar approach to the segmentation problem, and discuss some possible applications of our technique.

We shall refer to objects in Liu and Lawrence (1999) adding letter "L" after the formula and figure numbers.

The important step in the algorithm of Liu and Lawrence is the segmentation to the given number of segments k. They consider all segmentations with k segments as statistically independent ("model independence assumption") and equally likely. An additional prior probability is introduced for the number of segments; this probability is inversely proportional to the number of ways to segment the sequence into the given number of blocks. The probabilities of all possible segmentations are summed using the model independence assumption to obtain the marginal likelihood for the whole model (note that our marginal likelihood (8) is calculated for a single segment). Thus, in contrast to our technique, the summands corresponding to the segmentations with different k participate in the marginal likelihood sum with different weights.

In our approach, the value that is the closest relative of the marginal likelihood of the Liu and Lawrence model is the partition function (1). The difference is that in the technique used by Liu and Lawrence the terms corresponding to the segmentations with different k are weighted to the number of ways to segment the sequence into k parts.

The first step in the Liu and Lawrence technique is the calculation of the posterior probability of observing the segmentation of the sequence into k segments, which would in our terms correspond to the calculation of the partition function for all possible partitions into k segments. They do it using a recursive algorithm, which requires in our estimates about $N^2 \cdot k/2$ steps, where N is the length of the whole sequence.

The resulting values allow them to obtain the posterior distribution of the number of segments in the sequence. In our terms, the posterior distributions for the segmentation into k segments is proportional to the partition function of the segmentations into k segments related to the number of its summands, or the mean likelihood of the segmentation. This is a useful value, allowing one to estimate the smallest number of segments necessary to make a reasonable segmentation. When k is smaller than the "real number" of segments in the sequence, then all possible segments participating in the k-segmentation are "bad," i.e., overlap with several "natural" segments. (This explains the steep front of the histogram in Figure 4L.) The smooth tail of the distribution appears to be less informative. This agrees with the remark of Stephens (1994), who also considered this problem, that the number of models with different k should be small, and it is generally very difficult to make a formal choice between the models. As we do not consider independent segmentations with different k, we do not calculate this value.

The next value calculated by Liu and Lawrence is the probability that a change point will occur at position v (19L). Note that probabilities in (19L) are summed without their combinatorial weights, thus they coincide with our partition functions ZL and ZR, differing only in the normalization.

Using our version of dynamic programming we calculate the numerator much faster (for about $k^2 \cdot (N-k)^2$ instead of $k^3 \cdot (N-k)^3$).

Moreover, we have observed that probability (2) is of limited use for complicated sequences. The problem is that expression (2) has high values at all points close to the "true" change point (note the bell-like maxima in Figure 5L). The "true" change points definitely correspond to the maxima, but in the case of a complex sequence, a computer search for the true maxima (excluding probably the several highest ones) is rather difficult. That is why we use the partition function to filter the boundaries already produced by the optimal segmentation: the change points between optimal blocks (which as a rule are very short, see Figures 1, 2) are always placed at correct positions as are the change points between the longer blocks after filtration.

Liu and Lawrence also obtained several values, such as distribution of boundaries and distribution of the segmented composition using a sampling procedure. These are important results (especially those related to distribution of the composition). However, it is not clear whether it is possible to obtain reliable sampling results for a sequence of several thousand base pairs. Since we are more interested in the specific positions of change points rather than in the distributions, we did not use sampling.

We are going now to discuss the possible applications of our segmentation agorithm in biological studies. We believe that it may be applied to every problem in which the nucleic acid composition is important. For instance, one can mention the compositional difference between exons and introns, various dependences in DNA, and RNA composition associated with structural motifs, e.g., loops. Moreover, long eukaryotic genomic regions have been sequenced, so it will be possible to study the isochoric organization of eukaryotic genes directly. Among other examples of biological problems that can be solved with our technique, we would like to mention the search of gene migration events between genomes with highly different average nucleotide content.

Our technique also may be helpful to calculate the DNA melting curve in thermodynamic experiments.

The segmentation obtained can be used for further studies on DNA complexity or fractal structure (Karlin and Brendel, 1993; Peng *et al.*, 1992). To this end, the distribution of the lengths of homogeneous blocks may be very informative. Obviously, some, if not a decisive, contribution to the sequence complexity measured with various informational measures is explained by the compositional patchiness of the DNA sequence. This patchiness is mostly associated with sense or structural units, exons for instance (Herzel and Grosse, 1997). However, our preliminary studies (Ramensky *et al.*, 1998) indicate that long homogeneous segments are not necessarily associated with exons and introns and are likely to be of a much broader origin. A direct study of the DNA patchiness at the level of full genomes will allow one to understand to what extent the complexity of DNA is due to compositional patterns, and to what extent it is due to more subtle correlations.

The investigation of the length distribution of homogeneous segments obtained for long genomic regions will provide a final answer in the discussion on the origin of the fractal dependences in DNA correlation plots (Karlin and Brendel, 1993).

APPENDIX

Marginal and extreme likelihood in the case of large N

Using the Stirling approximation $n! = \sqrt{2\pi n} n^n e^{-n}$, it is easy to calculate the form of extreme likelihood and marginal likelihood for large N. Consider the following fraction:

$$\frac{(N+L-1)!}{n_1!\ldots n_L!} = \frac{(n_1+\ldots+n_L+L-1)!}{n_1!\ldots n_L!}. \tag{A1}$$

The Stirling approximation of the numerator contains the term

$$(N + L - 1)^{N+L-1} = N^N \left(1 + \frac{L-1}{N}\right)^N (N + L - 1)^{L-1}. \tag{A2}$$

The second multiplier equals e^{L-1} for large N. Then, the direct Stirling approximation of the fraction (A1) is

$$\frac{(N + L - 1)!}{n_1! \dots n_L!} \to \sqrt{\frac{N^L}{n_1 \dots n_L}} \left(\frac{N}{2\pi}\right)^{(L-1)/2} \left(\frac{N}{n_1}\right)^{n_1} \dots \left(\frac{N}{n_L}\right)^{n_L}. \tag{A3}$$

Introducing the point estimator $p_{\mathrm{k}} = n_{\mathrm{k}}/N$, one can rewrite (A3) as

$$\frac{(N + L - 1)!}{n_1! \dots n_L!} \to \left(\frac{N}{2\pi}\right)^{(L-1)/2} p_1^{-(n_1+1/2)} \dots p_L^{-(n_L+1/2)}. \tag{A4}$$

Hence, it is easy to calculate the approximations for the extreme likelihood

$$\frac{(N + L - 1)!}{n_1! \dots n_L!} \frac{(2n_1)! \dots (2n_L)!}{(2N + L - 1)!} \to 2^{-(L-1)/2} p_1^{n_1} \dots p_L^{n_L}, \tag{A5}$$

and marginal likelihood.

ACKNOWLEDGMENTS

The authors wish to thank A.M. Leontovich and P.S. Lifanov for their help in the early stage of this work, S.R. Sunyaev for useful comments and attention, M.S. Gelfand for his constant interest and for discussion on biological applications, D. Haussler for stimulating discussion and for the reference to the work of Liu and Lawrence (1998), and R. Guigo for discussion and for the copy of the proceedings of the Halkidiki (1997) workshop (Lawrence, 1997). The authors thank the anonymous referee for many valuable comments.

This study has been partially supported by Russian Human Genome Program 18/48 (V.E.R., V.Ju.M., V.G.T) and 99-0153-F(018) (M.A.R.), RFBR grant N 980448755 and MGRI project 244. M.A.R. is also supported by an RFBR Project.

REFERENCES

Bains, W. 1993. Local self-similarity of sequence in mammalian nuclear DNA is modulated by a 180bp periodicity. *J. Theor. Biol.* 161, 137–143.

Bernaola-Galvan, P., Roman-Roldan, R., and Oliver, J.L. 1996. Compositional segmentation and long-range fractal correlation in DNA sequences. *Phys. Rev. E.* 53(5), 5181–5189.

Bernardi, G. 1989. The isochore organization of the human genome. *Ann. Rev. of Gen.* 23, 637–661.

Bernardi, G. 1995. The human genome: organization and evolutionary history. *Ann. Rev. Gen.* 29, 445–476.

Bernardi, G., Olofson, B., Filipski. J., Zerial. M., Salinas, J. *et al.* 1985. The mosaic genome of warm-blooded vertebrate. *Science*, 228, 953–958.

Booth, N.B., and Smith A.F.M. 1992. A Bayesian approach to identification of change-points. *J. Econometr.* 19, 7–22.

Chechetkin, V.R., and Lobzin, V.V. 1998. Study of correlations in segmented DNA sequences: application to structural coupling between exons and introns. 190, 69–83.

Chernoff, H., and Zacks, S. 1964. Estimating the correct mean of a normal distribution which is subjected to a change in time. *Am. Math. Statist.*, 35, 999–1018.

D'Onofrio, G., Mouchiroud D., Aissani B., Gautier C., and Bernardi G. 1991. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J. Mol. Evol.* 32, 504–510.

Dunbrack, R.L., and Cohen, F.B. 1997. Bayesian statistical analysis of protein side-chain rotamer preferences, *Protein Science*, 6, 1661–1681.

Fickett, J.W. 1982. Recognition of protein coding regions in DNA sequences. *Nucl. Acid. Res.* 10, 5303–5318.

Finkelstein, A.V., and Roytberg, M.A. 1993. Computation of biopolymers: A general approach to different problems. *BioSystems*, 30, 1–19.

Fomin, V.N. 1984. *Rekurrentnoe otsenivanie i adaptivnaya fil'tratsiya* (Recurrent Inference and Adaptive Filtration). Moscow, Nauka.

Frank, G.K., and Makeev, V.Ju. 1997. G and T nucleotide contents show specie-invariant negative correlation for all three codon positions. *J. Biomol. Str. Dynam.* 14, 629–639.

Gelfand, M.S. 1992. Statistical analysis and prediction of the exonic structure of human genes. *J. Mol. Evol.* 35, 239–253.

Gelfand, M.S. 1995. Prediction of function in DNA sequence analysis *J. Comp. Biol.* 2, 87–117.

Gelfand, M.S., and Koonin, E.V. 1997. Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. *Nucl. Acid. Res.* 27, 2430–2439.

Grosse, I. 1996. Estimating entropies from finite samples. *Dynamik—Evolution—Strukturen*, J. Freund ed., Kosster Verlag, Berlin, 181–190.

Guigo, R., and Fickett, J.W. 1995. Distinctive sequence features in protein coding, genic non-coding and intergenic human DNA, *J. Mol. Biol.* 253, 51–60.

Herzel, H., and Grosse, I. 1997. Correaltions in DNA sequences: The role of protein coding segments. *Phys. Rev. E.* 55, 800–810.

Karlin, S., and Brendel, V. 1993. Patchiness and correlations in DNA sequences. *Science*, 259, 677–680.

Krogh, A., Brown, M., Mian, S., Sjolander, K., and Haussler, D 1994(a). Protein modeling using hidden Markov models. *J. Mol. Biol.* 235, 1501–1531.

Krogh, A., Mian, I.S., and Haussler, D. 1994(b). A hidden Markov model that finds genes in *E. coli* DNA. *Nucl. Acid. Res.* 22, 4768–4778.

Kypr, J., and Mrazek, J. 1986. Occurrence of nucleotide triplets in genes and secondary structure of the coded proteins. *Int. J. Biol. Macromol.* 9(2), 49–53.

Kypr, J., and Mrazek, J. 1995. Middle-range clustering of nucleotides in genomes. *Comp. Appl. Bio. Sci.* 11, 195–199.

Kypr, J., Mrazek, J., and Reich, J. 1989. Nucleotide composition bias and CpG dinucleotide content in the genomes of HIV and HTLV 1/2. *Biochem. Biophys. Acta.* 1009, 280–282.

Lawrence, C.E. 1997. Bayesian nioinformatics *5th International Conference on Intelligent Systems for Molecular Biology, Halkidiki-Greece*, 1997, 24.

Li, W., and Kaneko E. 1992. DNA correlations (scientific correspondence) *Nature*, 360, 635–636.

Li, W. 1994. Understanding long-range correlations in DNA sequences. *Physica D.* 75, 392–416.

Li, W. 1997. The study of correlation structure of DNA sequences: A critical review. *Computer and Chemistry* 21(4), 257–295.

Lin, J. 1991. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theor.* 37, 145–149.

Liu, S.L., and Lawrence, C.E. 1998. Bayesian inference on biopolymer models, *Stanford Statistical Department Technical Report* (to be published in Journal of Am. Stat. Ass). 24.

Liu, S.L., and Lawrence, C.E. 1999 Bayesian inference on biopolymer models, *Bioinformatics* 15, 38–52.

Mrazek, J., and Kypr, J. 1994. Biased distribution of adenine and thymine in gene nucleotide sequences. *J. Mol. Evol.* 39, 439–447.

Ossadnik, S.M., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Peng, C.-K., Simons, M., and Stanley, H.E. 1994. Correlation approach to identify coding regions in DNA sequences. *Biophysical Journal*, 67, 64–70.

Peng, C.-K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M., and Stanley, H.E. 1992. Long-range correlations in nucleotide sequences. *Nature*, 356, 168–170.

Ramensky, V.E., Makeev, V.Ju., Roitberg, M.A., and Tumanyan, V.G. 1998. Bayesian approach to DNA segmentation. In: *Proceedings of the Theoretical Approach in Biophysics conference.* Moscow. Moscow State University.

Reese, M.G., Eeckman, F.H., Kulp, D., and Haussler, D. 1997. Improved splice site detection in Genie. *J. Comp. Biol.* 4:3, 311–323.

Roman-Roldan, R., Bernaola-Galvan, P., and Oliver, J.L. 1998. Sequence compositional complexity of DNA through an entropic segmentation method. *Phys. Rev. Lett.* 80(4), 1344–1347.

Rozanov, Yu.M. 1985. *Teoriya veroyatnosti, sluchainye processy i matematicheskaya statistika (Probability Theory, Stochastic Processes and Mathematical Statistics)*, Moscow, Nauka.

Ryan, M.S., and Nudd, G.R. 1993. The Viterbi algorithm. *Warwick Research Report RR238* available via e-mail from msr@dcs.warwick.ac.uk.

Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I.S., and Haussler, D. 1996. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.* 12, 327–345.

Smith, A.F.M. 1975. A Bayesian approach to inference about a changepoint in a sequence of random variables. *Biometrika.* 67, 79–84.

Stephens, D.A. 1994. Bayesian retrospective multiple-changepoint identification. *Appl. Stat.* 43, 159–178.

Sueoka, N. 1959. A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation. *Proc. Natl. Acad. Sci.* 45, 1480–1490.

Trifonov, E.N., and Sussman, J.L. 1980. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl. Acad. Sci.* 77(7), 3816–3820.

Tsonis, A.A., Elsner, J.B., and Panagiotis, A. 1991. Periodicity in DNA coding sequences: implication in evolution. *J. Theor. Biol.* 151, 323–331.

Wolf, D.R., and Wolpert, D.H. 1993. Estimating functions of probability distributions from a finite set of samples. Part I and II. *Los Alamos National Laboratory Report No. LA-UR-93-833* (unpublished). Send emails to `comp-gas@xyz.lanl.gov` with subjects "get 9403001" and "get 9403002" to get an encoded postscript version.

Wolpert, D.H., and Wolf, D.R. 1995. Estimating functions of probability distributions from a finite set of samples. *Phys. Rew. E.* 52, 6841–6854.

Zazzi, H., Nikosjkov, A., Hall, K., and Luthman, H. 1998. Structural and transcription regulation of the human insulin-like growth factor binding protein 4 gene (IGFBP4) *Genomics*, 49, 401–410

Address correspondence to:
V.E. Ramensky
Engelhardt Institute of Molecular Biology
Vavilova, 32 Moscow
117984 Russia

E-mail: ramensky@imb.ac.ru